图书情报工作
第62卷 第11期 2018年6月

借阅场景下图书专业性质量测度方法和图书个性化推荐服务方法*

■ 李树青[1] 庄光光[1] 秦嘉杭[2] 徐侠[3]

[1] 南京财经大学信息工程学院 南京 210046 [2] 南京财经大学图书馆 南京 210046

[3] 南京邮电大学管理学院 南京 210046

摘要：[目的/意义]以现有图书馆借阅记录为基础，结合图书阅读相关性进行深入挖掘，探讨识别借阅场景下图书专业性质量和实现相应个性化图书推荐服务的有效方法。[方法/过程]利用图书的阅读相关性提出图书相关性链接关系，结合图书质量的迭代识别算法来识别专业图书资源。同时利用图书类别相关性链接关系，提出读者用户个性化模式的表达方法，并从长期兴趣推荐和短期兴趣的即时推荐两个方面给出个性化图书推荐策略设计原理和实现方法。[结果/结论]在图书质量识别方面，该方法更易于识别出专业性较强的优质图书资源，适用面比较灵活，也可以在限定图书范围内进行专业图书识别。在个性化图书推荐方面，发现不论长期兴趣推荐方法还是短期兴趣推荐方法，第二类用户的平均推荐命中度要高于第一类用户，在第一类用户中，最高相似度区间(75%以上)和较低相似度区间(15% －50%)的短期兴趣推荐方法的平均推荐命中度要高于长期兴趣推荐方法。本研究通过读者借阅序列分析方法识别专业图书并实现相应的个性化推荐图书方法，有利于改善现有图书馆借阅服务水平和提高读者的满意度。

关键词：个性化推荐 图书借阅 图书馆服务 图书质量

分类号：G202

DOI：10.13266/j.issn.0252 －3116.2018.11.006

1 引言

个性化推荐服务目前最为主要的应用在于电子商务领域，因为它的成功效应，也逐渐在很多其他领域不断被应用。但是不同于电子商务领域，图书馆个性化推荐服务所能利用的读者行为信息具有自己的特点，比如能反映读者用户兴趣程度的数据规模和种类比较有限，图书馆借阅活动需要人员的到场，因此数据产生的规模相对有限，比如根据我们对南京财经大学2011到2014年共计4年的读者借阅统计数据分析发现，年平均借阅量约在3万次左右，人均每次借阅册数约为11本。在这相对有限的数据中，还缺乏类似于电子商务网站用户对所购商品的评分信息，而这个信息对于最为经典的协同过滤个性化推荐方法非常重要。但是从另一方面看，由于读者阅读内容和知识学习的自身规律，同一读者会对同一本图书或者同一类图书反复借阅，甚至会形成一种较为稳定的借阅轨迹，因此结合图书类别和时序关系可以得到更多的研究角度。这些都说明，对于面向图书馆借阅服务的个性化推荐方法，相关设计思想和实现方法都需要进行专门研究。

从推荐流程来看，要想实现更为良好的服务效果，有两个关键环节需要深入的研究：①给当前读者用户推荐的图书应该避免马太效应，一般图书推荐都会采用诸如热门图书和借阅量排行等流行度指标，而在借阅场景下，读者的需求具有多样化特征，同时常常存在着对专业优质图书资源的现实需求，因此设计有效的优质专业性图书识别方法是完善图书馆个性化推荐服务整体流程的关键内容；②对当前读者用户兴趣特征的有效表达，这方面可以通过分析读者已有借阅历史来得到，实验证明进一步结合借阅历史及其读者关系可以提供更为有效的表达途径[1]。这两部分分别构成

* 本文系2016年国家社会科学基金项目"基于大数据分析的数字图书馆个性化服务模式创新研究"(项目编号:16BTQ030)研究成果之一。
作者简介：李树青(ORCID:0000 －0001 －9814 －5766)，教授，博士，E-mail:leeshuqing@163.com；庄光光，硕士研究生；秦嘉杭，馆长，博士；徐侠，副教授，博士。
收稿日期:2017 －11 －07 修回日期:2018 －02 －08 本文起止页码:53 －63 本文责任编辑:易飞

了本文的两个主要研究内容。

2 文献回顾

图书馆个性化推荐服务可以极大地改善现有图书馆借阅服务的读者用户体验,提高图书馆现有图书的综合利用率[2]。相关实现方法有很多,随着对Web2.0和媒体受众影响力的相关研究不断深入,借助媒体受众更多地参与信息产品的创造、传播和分享,并根据媒体受众的行为来间接发现信息资源的价值和识别优质信息资源也成为一个极具潜力的研究方法。基于现有的图书借阅记录,可以进行有效的图书质量分析方法研究,也可以提供有效的推荐服务策略。近年来相关研究逐渐增多。

对于图书质量的测度研究而言,优质专业性图书有多种识别标准,既可以通过专家评价,也可以通过读者评价,甚至还可以通过其他一些间接方法来进行。对于图书馆借阅服务而言,优质专业性图书的判定必须从内容和读者两方面来综合考虑。高质量专业性图书内容本身的强专业性特点往往使得该类图书在特定借阅读者群体中并不会受到广泛关注,也难以获得整体读者群体的认可,反之,广泛受到读者关注的图书往往是因为读者群体读书意愿多样化的影响而非因为图书内容本身的高质量专业性,图书的质量和专业性难以直接通过借阅量得以反映。因此,借助间接方法综合考虑上述两个方面的影响,可以提供更为全面的优质专业性图书识别方法。

如学者利用用户评价和评分实现的协同过滤方法,提出对专业性图书质量的测度方法,并据此实现个性化图书推荐服务[3]。还有学者指出借阅次数和平均借阅次数等传统指标存在很多弊端,如即使单本图书借阅次数再可观也不能说明需要大量购买同一本图书,同时图书还具有时效性,专家推荐和读者提交等人工方法又存在主观性强和不易于推广使用的问题,但针对这些问题所提出的具体图书质量评价方法上只使用了图书N指数方法,该方法只根据图书类别来进行分析,没有实现对具体每本图书的质量评估[4]。另有学者提出利用图书平均每次被外借的时间、被外借次数以及是否是新书等指标设计评估方法,以衡量图书的受欢迎程度,间接表达图书的质量[5]。

利用图书借阅记录的更多相关研究最终目的还是服务于有效的推荐方法设计。如有学者利用大数据资源和关联规则分析方法来从读者借阅记录中发现读者兴趣模式,通过改进的频繁模式增长算法,并据此实现线上和线下的个性化推荐服务[6]。还有学者借鉴电子商务推荐系统冷启动处理办法,利用改进的K-medoids算法对已有读者、已有图书进行基于借还时间间隔的聚类,实现了面向新读者和新图书的数字图书馆个性化推荐服务[7]。

值得注意的是,这些方法极大依赖于对读者借阅图书记录的有效处理和相关读者借阅行为的准确理解。目前具有的个性化图书推荐相关研究方法有很多,如关联规则方法[8-9]、主题模型方法[10]等。单纯使用借阅者的历史借阅数据的传统图书推荐算法通常会造成推荐的精准度偏低,因此结合基于图书之间的相似度和借阅者之间相似度的综合方法[11],和结合诸如图书分类的多特征方法[12-13]都在图书推荐研究中取得了较好的成效。

对于读者借阅行为本身而言,有学者将读者的借阅行为分为4种不同的类型,即续借、超期借阅(长期超借与短期超借)、正常借阅、盲目借阅,分别计算其相对借阅时间,并认为盲目借阅和长期超借都不能有效反映读者兴趣[14]。还有学者将读者的借阅行为分为借阅、续借、预约3种不同的类型,并认为借阅产生的时间与读者的兴趣度存在联系,因此通过引入时间衰减策略来完善读者借阅行为的分析方法[15]。

近年来,有越来越多的学者着重从借阅信息形成的借阅记录网络结构进行分析,从而给相关图书推荐服务提供了一个新的研究起点。如有学者从网络结构角度进行过较为全面的分析,指出借阅网络具有相对较高的平均集聚系数、较小的平均最短路径长度,具有明显的复杂网络结构特征[16-17]。还有学者利用借阅网络从中得到读者的共同借阅关系,通过向网络添加用户个人属性和图书分类,进一步研究不同类别的用户和不同类别的图书之间的借阅联系强度,从而提出很多有针对性的个性化图书馆借阅服务建议措施[18]。

在这些研究方法中,利用借阅二分网络结构的分析方法是一种较为常见的方法[19]。我们在前期的实验中,利用读者借阅行为特征来形成判断图书可推荐质量的依据,并结合借阅二分网络结构设计了一种测度图书可推荐质量的迭代算法,提出了包括特定主题的图书推荐服务、现有所借图书的修正型推荐服务和新书推荐服务3种个性化图书推荐服务形式[20]。类似的研究还有很多,如有学者通过能量传递六步算法反复在借阅二分网络结构中扩散权值,从而获得由不同权重图书组成的推荐列表,实现个性化图书借阅推送服务[21]。更有学者进一步结合从Web网络上抓取

的图书购买记录，结合高校图书馆借阅记录，利用用户对图书的评分作为借阅二分网络图的权值，综合用户之间对不同图书评分的偏好预测，实现对相同图书评分的偏好预测和借阅偏好预测，进而完成个性化图书推荐服务[22]。

进一步从推荐策略创新的角度来看，利用时间及其演化信息来增强用户个性化兴趣模式的识别能力和用户兴趣特征信息的表达能力，构成了个性化推荐服务研究领域一个富有潜力的研究方向[23]。然而，结合时间因素的个性化图书推荐研究仍不多见，有学者提出只利用最近一学期借阅记录计算用户短期需求偏好，及利用用户的整个借阅记录计算用户长期需求偏好的方法[24]。笔者前期的研究也逐渐发现结合时间信息分析的重要性，并取得了一些初步研究成果，如利用加权兴趣表达方法提出了加权关键词共现时间元，通过对关键词时序路径的发现和对关键词时序网络结构的表达，对个性化学术研究时序路径的发现方法及其可视化界面设计进行了研究[25]。本文将继续对此进行研究探索，并且力图实现对长期兴趣推荐方法和短期兴趣推荐方法的比较和特点分析。

从国内高校图书馆的应用实际来看，个性化图书推荐服务的普及依然进展有限，相关网络应用和服务都十分欠缺，除南京大学等部分高校外，很少有大学的图书馆系统提供此类的个性化图书推荐服务，即使提供推荐服务也非个性化推荐服务[24]。这也充分说明了相关研究的必要性。本文即以读者借阅记录二分网络结构分析入手，完成对优质专业性图书的识别和读者用户相似度的测度，从而探索一种新的个性化专业图书推荐服务方法。

3 优质专业图书识别方法

3.1 基本思路说明

对于图书馆借阅服务而言，读者借阅记录能够反映读者自己对于图书的阅读意愿和关注度。专业读者通常更能理解所借专业图书的质量，图书借阅记录本身就能体现借阅者对相关图书的一种认可。通常越是优秀的专业读者越能借阅到更为优质的专业图书。由于专业性的差异，图书馆测度优质图书一定不能忽略专业读者的认可程度。在读者借阅记录中，连续的借阅记录往往能够表明读者对于所借阅的一系列图书的关注程度，也能反映图书之间的关联程度。

然而，测定专业读者的质量和辨析优秀程度并不容易，单纯的图书借阅记录也很难提供直接的分析依据。因此，更为常见的方法是从被借图书本身入手，通过分析被借图书之间的联系，来测度图书之间的关联性和区分图书的质量。

3.2 图书的阅读相关性

在读者借阅图书的历史记录中，一般能形成如表1所示格式的借阅信息：

表1 包括时间序列信息的读者借阅图书的历史记录

读者ID	书籍ID	借阅时间	归还时间	借阅时长（天数）	借阅序列号
023	118808	2011－03－30	2011－09－05	159	1
023	153604	2011－04－02	2011－05－20	48	2
023	49490	2011－04－06	2011－09－05	152	3
023	160691	2011－05－20	2011－09－05	108	4
023	147269	2011－05－20	2011－10－07	140	4
023	153604	2011－09－28	2011－11－30	63	5

得到读者借阅序列的方法有很多种，关键是如何提取所需的时间信息，不同时间信息提取方式会得到不同的序列生成方法。借阅时间和归还时间是最为基础的时间信息，通常读者会在一次借阅操作中完成对多本图书的集中借阅，同样读者也会在一次归还操作中完成对多本图书的集中归还，从而形成一种借还顺序交叉的借阅序列。

基于目前的分析，我们提出如下假设：

假设：对于全部读者而言，所借图书A如果能经常在借阅过图书B后，并且尚未归还图书B前被借阅，这说明图书A和图书B具有一定的阅读相关性。

这种阅读相关性具有多种语义解释的可能，既可能说明图书之间存在明显的内容相关性，也可能说明具有外延的阅读扩展关系，当然也可能存在着其他未知原因，然而由于出现频次的数量较高，在很大程度上可以提供一个测度图书关联度的有效途径，即使对于背后原因不能确定，也可以据此提供一个值得深入挖掘的数据资源体。为了进一步分析，需要给出量化测度这种阅读相关性的方法。

借阅时长是较为直观的数据内容，然而我们在实验中也发现，单纯利用天数等借阅时长单位进行测度往往带来很多不利的影响。主要原因有很多，比如受假期影响，如表1所示，就存在明显的暑假影响特征，再如个人阅读习惯的影响，不少读者平均借阅时长较长，并不表示他们对于这些图书都感兴趣，有时甚至恰恰相反，反映了读者借阅活跃程度较低。因此在我们的测度方法设计中，对借阅时长没有考虑。

借阅序列号是指一个用户在所有的借阅记录中每

一次借阅操作的唯一标识号，也就是说，在用户第 1 次借阅操作时，所有借阅记录的借阅序列号都被分配为 1，以后每次借阅都递增，因为读者可能一次借阅多本图书，所以部分借阅记录中存在同一借阅序列号对应多本借阅图书的情况。通过该指标可以表达不同图书的借阅次序。

3.3 图书相关性链接关系的构建方法

按照图书阅读相关性的定义，我们从读者借阅序列中抽取出所需的图书相关性链接。如表 1 都为同一读者的借阅记录，按照借阅序列号的递增关系和借阅时间的包含关系，可以得到如图 1 所示的图书相关性链接关系：

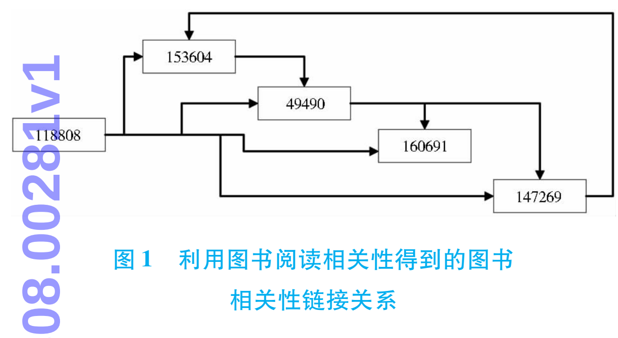


图 1 利用图书阅读相关性得到的图书相关性链接关系

由图 1 可以看出，该读者在归还 118808 图书前，后续又借阅了 4 本图书，分别是 153604、49490、160691、147269，因此相关节点之间建立了有向连接。而 153604 在被归还前只存在一本后续借阅图书 49490，因此只有一条有向连接。

如果对所有读者的全部借阅记录进行同样的处理，得到的这种图书相关性链接图应该是一个典型的网络结构，从实验分析来看，它也具有复杂网络的典型特点。值得说明的是，图书相关性链接图有很多生成方法，传统方法往往以共同借阅关系来构造，本文所提方法可以避免传统方法中借阅量较高图书往往具有较高链入节点的特点，相反，只有具有较多图书阅读相关性的图书才能形成较高的链入节点或者链出节点，这显然给我们提供了一个新颖的图书质量测度依据和用户相似度测度依据。

从读者用户行为的角度来分析，我们认为读者在连续借阅行为中，随着不断地了解所借图书的内容，更易于在后续的借阅行为中借阅到与图书内容相关的更为专业的图书资源。因此，该方法所测度出来的优质图书往往在内容上更为专业，从而为个性化推荐服务提供了良好的推荐客体资源。

3.4 图书专业性质量的迭代识别算法

在复杂网络结构中使用迭代计算方法可以得到权值收敛后的节点信息，如 PageRank 方法等[26]。该方法主要在由众多节点组成的网络结构中，通过给每个节点赋予初始权值，然后利用基于链出关系的权值扩展方法，通过多次迭代计算最终得到稳定的节点权值，从而实现对网络节点质量的有效测度。

然而这些传统方法在分配节点权值和权值扩散策略选择上，并没有考虑到特定应用领域中的特点。前文指出，图书阅读相关性是建立在具有一定出现频次的这个重要前提上，不管是对于单一读者而言，还是对于所有读者而言，经常出现的图书相关性链接关系更能说明相关链出和链入图书节点的阅读相关性。按照这个设计原则，需要对传统网络节点迭代算法进行必要的修正。

在标准 PageRank 方法的基础上，本文提出了如式(1)所示的算法思路：

$$WeightRPR(Book_i) = c\frac{\sum_{DOC_{ij} \in F(Book_i)} WeightRPR(Book_{ij}) \times count_{ij}}{N_{F(Book_i)}} + (1-c) \quad 式(1)$$

WeightRPR 表示基于修正 PageRank 方法的图书权值，F(Booki)集合表示图书 Booki 在图书相关性链接关系图中所有链入图书的集合，$N_{F(Booki)}$ 表示该链入图书集合的图书数量。其中有 3 处需要说明：

(1) $Count_{i,j}$ 表示对应图书 $Book_i$ 和 $Book_j$ 之间链接关系的数量，通过该系数放大每个链入图书权值的影响程度，据此反映经常出现的图书相关性链接关系更能说明相关链出和链入图书节点的阅读相关性。

(2) 在实际计算中，$Book_i$ 和 $Book_j$ 并非一定是不同的图书。相反，这种连续借阅同一图书的行为也更能反映该图书的质量，保留此类链接关系有助于发掘高质量的图书。

(3) 式(1)提供了一个基础的计算方法，在实际应用中，可以根据应用环境和需求做出更多调整，如将图书换成图书类别，据此来挖掘不同的图书类别在指定读者群体中的被关注程度，再如在特定读者群中使用该方法，可以分层了解不同读者群的关注点和最受关注的高质量专业性图书。

4 基于读者借阅序列分析的个性化专业性图书推荐方法

4.1 基本思路

读者借阅序列也可以给个性化图书推荐服务提供分析条件，具体可以形成多种表达用户个性化模式的方法和相关用户相似度判定方法，如直接利用借阅图

书为处理单元等。然而这些方法往往都存在一定的稀疏性问题和表达精度问题。

要想回答借阅过同样图书的用户是否为兴趣相似的用户,我们可以先回答另外一个相关问题:是否兴趣相似的用户会借阅同样的图书? 以下列举几种常见的情况加以分析:

(1)在图书馆借阅系统中,很多图书具有不同的版本,而且往往同一版本的图书具有多个不同的 ID 号,即使是同一本图书,由于借阅册数限制,也决定了不可能让所有想借阅的读者都能借阅到,因此读者往往只能选择相似图书或者类似的其他版本。

(2)借阅行为的动机比较复杂,所借图书对于读者兴趣的反映程度也并非直接对应,很多情况下,读者存在着试探性的、偶发性的借阅行为,即使是与专业相关的图书借阅,也存在自己理解的差异,不同的读者对于同一本图书也会有着不同的评价标准,所借图书反映读者自身兴趣的能力也各有不同。

从上述分析可以看出,直接以借阅过同样图书作为判断用户兴趣相似的依据存在着自身的局限性。以读者借阅序列中存在着的借阅次序为基础,可以从下面两个方面来改进:

(1)使用图书类别作为处理单位,该方法可以较好地在读者兴趣准确还原和避免表达过于细致带来的漏报这两个方面间进行折中处理。具体方法可以利用每本图书都标注的《中国图书分类法》(简称《中图法》)中的分类号。由于该分类号存在多级目录层次,因此我们在实际实验中采用了只保留分类号前面英文字母前缀和后 2 位数字的处理策略,如对于“F752.68/27”,保留结果为“F75”,对于“TP391.13/24”,保留结果为“TP39”等。

(2)以前文所述的图书相关性链接关系为基础,将图书链接关系映射为对应的图书类别映射关系,再以图书类别相关性链接关系作为兴趣表达单元,如表 2 所示:

表 2　图书类别相关性链接关系的例子

读者 ID	链入图书类别	链出图书类别	图书类别相关性链接	频次
2120120324	I25	I53	I25——I53	1
2120120324	I20	I20	I20——I20	7
2120120324	I56	H31	I56——H31	15
10199359	I31	I26	I31——I26	4
10199359	O14	O14	O14——O14	11
10199359	O13	I53	O13——I53	36
10199359	R22	O21	R22——O21	1

其中的频次越大,有效性越明显,该信息也给后续的用户相似度计算提供了量化的数据基础。

4.2　相似度计算和推荐方法设计

我们以所有图书类别相关性链接为向量单元,以频次信息作为向量单元值,就可以得到每个读者用户的兴趣特征向量。如对于每个读者用户 i,都可以得到如式(2)所示的读者用户兴趣模式:

$$ReaderVector_i = \{(\text{图书类别相关性链接 }1,\text{频次 }1),(\text{图书类别相关性链接 }2,\text{频次 }2),\cdots\cdots,(\text{图书类别相关性链接 }m,\text{频次 }m)\}$$ 式(2)

其中 m 为图书类别相关性链接总数量。

具体的读者用户相似度计算方法可以采用皮尔逊系数或者余弦夹角系数等,最终可以得到每位用户与其他相关用户的相似度。由于读者数量和图书类别相关性链接数量众多,在实际计算中,可以通过设定读者用户具有相同图书类别相关性链接的数量阈值来限定比较范围。

本文首先对每个读者用户兴趣模式的权值进行规范化处理,为避免受单个用户频次绝对数量的影响,采用每个读者用户最大频次去除其向量每一个频次值的方法进行权值规范化处理。然后,对每两个读者用户兴趣模式向量采用余弦夹角系数得到最终的用户相似度:

$$sim(ReaderVector_i, ReaderVector_j) = \frac{ReaderVector_i \cdot ReaderVector_j}{|ReaderVector_i||ReaderVector_j|}$$ 式(3)

在个性化推荐环节上,首先对于目标读者用户,得到最为相似的其他读者用户序列,在实际计算中,可以设定相似度阈值来控制该序列的大小。同时,我们设计了两种具有不同服务目标的个性化图书推荐服务形式:

(1)长期兴趣推荐。根据目标用户所有的借阅情况,获取相关借阅图书的类别信息,据此再到最为相似的其他读者用户序列中,汇总得到推荐图书列表,并按照前文所述的优质图书识别标准,倒序输出推荐图书列表。该种推荐形式主要面向读者用户的长期兴趣特征,所推荐的内容具有一定的稳定性和用户关联性。

(2)短期兴趣的即时推荐。根据最近 n 次目标读者用户的借阅情况(n 可以根据实验数据情况选择,如 2 次或者 3 次等),获取相关借阅图书的类别信息,据此到最为相似的其他读者用户序列中,按照最近一次借阅情况,汇总得到推荐图书列表,并按照前文所述的

优质图书识别标准,倒序输出即时推荐图书列表。该种推荐形式主要面向读者用户的短期兴趣特征,所推荐的内容具有强时效性。

5 实验

5.1 实验环境准备

利用南京财经大学图书馆汇文借阅系统 2011 年 1 月 1 日至 2014 年 6 月 16 日近 4 年的图书借阅记录作为实验数据,其中得到的有效借阅记录数据量为 1 076 749条,涉及的图书为138 696种,每种图书都有一个唯一的图书 ID,读者为42 750位,包括着学校教师和近 4 届的本科生和研究生。

为了对比试验结果,我们保留了全部读者所有最近一次借阅的内容作为对比数据,没有用于优质图书识别实验和用户相似度计算实验,此类借阅记录都是没有归还日期的记录,总数为 13 791 条,占总借阅记录 1.28%。

5.2 专业性图书质量识别实验结果

按照前文所述图书相关性链接关系图构建方法,从实验数据集合中抽取了2 595 690条记录,涉及的图书总数为126 033,占全部图书总数的 90.87%。主要原因在于有部分读者借阅图书的频次很少而且间隔很长,而此类图书也较少被其他更多读者借阅,因此并非所有图书都能在图书相关性链接关系图中构成其他图书的链入节点或者链出节点。

下面为了方便显示识别效果,我们对检索结果给出对比展示,对比对象是按照现有图书馆借阅系统中以借阅量为倒序排列标准的常见输出结果,如表 3 – 表 5 所示:

表 3 信息检索类图书(《中图法》分类号为 G25)的查询结果对比

(a)根据图书相关性链接关系图迭代算法识别的专业性图书

书名	作者	借阅量(册)	WeightRPR 权值	当当评分
信息检索理论与技术	苏新宁,主编	6	5.373 1E – 2	87.50%
现代信息检索	B. RICARDO, R. BERTHIER 等,著	11	4.669 0E – 2	99.40%
信息检索	陈明兵,主编	1	3.195 1E – 2	NULL
信息检索原理与技术	夏立新, 金燕,方志等,编著	4	2.804 9E – 2	99.20%
专利信息检索与利用	阚元汉,主编	4	2.688 0E – 2	100%

(b)根据借阅量倒序排列方法识别的热门图书

书名	作者	借阅量(册)	WeightRPR 权值	当当评分
信息检索问题集萃与实用案例	曹志梅, 范亚芳, 蒲筱哥,编著	11	2.411 2E – 2	NULL
现代信息检索	B. RICARDO, R. BERTHIER 等,著	11	4.669 0E – 2	99.40%
信息检索导论	D. CHRISTOPHER, R. PRABHAKAR, S. HINRICH,著	10	8.781 1E – 3	100%
信息检索与分析利用.2 版	谢德林,主编	9	2.285 0E – 2	NULL
信息检索理论与技术	苏新宁,主编	6	5.373 1E – 2	87.50%

表 4 管理学原理类图书(《中图法》分类号为 C93)的查询结果对比

(a)根据图书相关性链接关系图迭代算法识别的专业性图书

书名	作者	借阅量(册)	WeightRPR 权值	当当评分
管理学原理	斯蒂芬·P·罗宾斯, 戴维·A·德森佐, 亨利·穆恩,著	47	0.182 5	99.10%
管理学	里基·W. 格里芬,著	24	0.156 2	98.60%
管理学(第 3 版)	杨文士,等,编著	17	0.151 4	100%
周三多《管理学》笔记和习题详解	金圣才,主编	34	0.148 7	98.40%
管理学(第 2 版)	周三多,主编	31	0.144 8	100%

(b)根据借阅量倒序排列方法识别的热门图书

书名	作者	借阅量(册)	WeightRPR 权值	当当评分
管理学原理	斯蒂芬·P·罗宾斯, 戴维·A·德森佐, 亨利·穆恩,著	47	1.825 4E – 1	99.10%
罗宾斯《管理学》(第 9 版) 笔记和课后习题(含考研真题) 详解	圣才学习网,主编	41	9.630 3E – 2	100%
罗宾斯《管理学》(第 9 版) 学习指导	史蒂文·考克斯, 阿雷萨·考克斯,著	38	1.077 6E – 1	100%
管理学精要	加里·戴斯勒,著	36	7.768 0E – 2	NULL
管理学习题与案例	姜仁良,主编	35	1.006 7E – 1	91.70%

表 5　数据挖掘类图书(标题含有“数据挖掘”)的查询结果对比

(a) 根据图书相关性链接关系图迭代算法识别的专业性图书

书名	作者	借阅量(册)	WeightRPR 权值	当当评分
数据挖掘:概念与技术	H. JIAWEI,K. MICHELINE,著	24	0.219 8	97.70%
数据挖掘导论	T. PANGNING,S. MICHAEL,K. VIPIN,著	25	0.181 4	80%
数据挖掘:实用机器学习技术	H. IAN,F. EIBE,著	19	0.145 3	99.50%
Excel 2007 数据挖掘完全手册	谢邦昌，朱建平，来升强,编著	24	0.129 1	100%
数据仓库与数据挖掘	廖开际,主编	24	0.122 5	98.40%

(b)根据借阅量倒序排列方法识别的热门图书

书名	作者	借阅量(册)	WeightRPR 权值	当当评分
数据挖掘导论	T. PANGNING,S. MICHAEL,K. VIPIN,著	25	0.181 4	80%
数据仓库与数据挖掘	廖开际,主编	24	0.122 5	98.40%
数据挖掘:概念与技术	JIAWEI H,MICHELINE K,著	24	0.219 8	97.70%
Excel 2007 数据挖掘完全手册	谢邦昌，朱建平，来升强,编著	24	0.129 2	100%
数据挖掘:实用机器学习技术	H. IAN,F. EIBE,著	19	0.145 4	80%

从表 3－表 5 中可以看出本文所述方法的几个特点:

(1)在图书质量的识别上,侧重于挖掘专业性较强的图书资源,这也是该方法的一个主要优势。事实上,在传统借阅量倒序排列方法识别的优质图书中,对于高校图书馆而言,一些考试类和试题类图书往往取得较高的借阅量。

比如在“信息检索”类中,根据借阅量倒序排列方法识别的优质图书排名第一的是案例分析类,而该本图书并不出现在根据图书相关性链接关系图迭代算法识别的专业图书目录中,在“管理学原理”类中,与学习笔记相关的有 3 本,排名分别为第二、第三和第五,而在根据图书相关性链接关系图迭代算法识别的专业图书目录中只有 1 本,排名只为第四。

再如数据挖掘类,根据图书相关性链接关系图迭代算法识别的专业图书目录中前三本都是经典的数据挖掘相关图书,其中第一本韩家炜(JIAWEIH)所著的图书更是数据挖掘领域最为著名的经典图书。然而这 3 本图书在根据借阅量倒序排列方法识别的优质图书不仅没有全部排在前列,而且韩家炜所著的经典图书甚至掉到排名第三。

(2)为了从定量的角度进行更为有效的对比评价,我们利用中文当当网对应图书的好评指标,根据 NDCG(Normalized Discounted Cumulative Gain)指标中认为一般情况下用户会优先点选排在前面的搜索结果这一基本思路,引入折算因子,并据此统计查询结果排名前五位的最终评分值,计算公式如下:

$$discountingScore = \sum_{i} Score_{i} * log(2)/log(1+i) \quad 式(4)$$

表 3 到表 5 三个对比结果的相关计算情况如表 6 所示:

表 6　本文算法和按照借阅量倒序方法的评分对比

例子	discountingScore 本文算法	discountingScore 借阅量倒序方法	提升度
表 3 例	2.316 228 128	1.465 640 381	58%
表 4 例	2.923 735 277	2.476 673 778	18%
表 5 例	2.7905 835 23	2.649 493 681	5%

我们随后完成了随机 20 个专业领域的图书查询测试,本文算法的 discountingScore 平均值为2.736 2,按照借阅量倒序方法的 discountingScore 平均值为 2.375 691 74,总体高 20.38%。同时,从结果来看,该方法适用于各个专业图书领域,不存在明显的专业差异,只受到图书查询结果数量的影响,即有些专业图书读者很少,无法有效地根据用户行为来进行更为有效的识别。

(3)该方法的适用面比较灵活,它可以在查询到的所有图书范围内去进行满足特定需求的分析,如表 3 和表 4 是在关键词查询和《中图法》分类号限定双重约束下获得的查询结果,而表 5 和表 6 则只是关键词查询结果,之所以采用不同的查询策略,主要原因在于图书标题和分类号在表征图书内容方面都存在各自的不足,如有些图书标题文字表达方法存在着较大的变化,如查询“管理学原理”,可能的相关图书标题却为“管理学基本原理”,甚至还有“公共管理学原理”这样的误判,而对于图书分类号而言更是如此,很多同一类型的图书都会因为作者和标注者理解的不同而放在不同的分类号中,如“数据挖掘”类常见的分类号有“TP274”“TP311”和“O212”等,这也是为什么表 5 和

表 6 没有采用分类号限定的原因。

5.3 读者相似度实验结果

实验中由于读者借阅记录内容的限制,并非所有读者都与其他读者具有相同的借阅图书类别,因此在借助于图书类别相关性链接得到的用户相似度结果中,实际得到的有效(相似度大于 0%)的读者总数为 23 937位,占全部读者用户比重 56%。

实验结果说明用户相似度数值具有较大的变化空间,从相似度为 100% 到0.000 006%,具体数值分布情况如表 7 所示:

表 7 用户相似度数值区间及其用户对数量

相似度区间	匹配用户对数量(单位:位)
100%	11
[90%,100%)	2 279
[80%,90%)	3 147
[70%,80%)	1 983
[60%,70%)	2 259
[50%,60%)	4 305
[40%,50%)	5 454
[30%,40%)	12 266
[20%,30%)	20 875
[10%,20%)	44 137
(0%,10%)	326 268

匹配用户对总数为 422 984 位,有近 77.13% 的用户对相似度低于 10%,在余下的 23% 的用户对中,涉及的读者用户总数为14 669位,占实际有效(相似度大于 0%)读者总数 61.3%。

5.4 个性化推荐实验结果

该实验是对个性化推荐方法本身效果进行验证,包括两个实验部分,分别测试长期兴趣推荐方法和短期兴趣的即时推荐方法。

实验所选择的测试读者用户对象主要分为两大类:一类是满足借阅记录量大于 200 条以上的读者用户,和每位受测读者用户进行匹配的相似读者用户至少满足借阅记录量大于 10 条以上。总共得到 139 位受测读者用户、25 220位进行匹配的相似读者,产生的读者用户匹配数量为3 505 580对;另一类是满足借阅量大于 100 条并且小于 200 条的读者用户,和每位受测读者用户进行匹配的相似读者用户至少满足借阅量大于 10 条以上。总共得到1 213 位受测读者用户、25 220位进行匹配的相似读者,产生的读者用户匹配数量为3 505 580对。

两个实验重点测试的是所推荐的图书类别是否为用户感兴趣的图书类别。因为保留了全部读者所有最近一次借阅的内容作为对比数据,所以我们将所有用户借阅行为中借阅的图书类别为图书类别相关性链接关系的链入,对于长期兴趣推荐方法而言,记录范围为每个读者用户的全部借阅记录,对于短期兴趣推荐方法而言,则只利用每个用户最多最近 3 次的全部借阅记录。然后,根据这些链入在推荐方法中观察推荐图书类别结果,并根据实际用户最近一次借阅的图书类别情况,统计推荐方法给出的图书类别在实际后续借阅行为中出现的比重,即推荐命中度,以此来测度个性化推荐方法本身的有效性,具体方法如式(5)所示:

$$推荐命中度=\frac{后续借阅出现推荐图书类别的用户总数}{所有用户总数} \quad 式(5)$$

对于第一类借阅量大于 200 条以上的读者用户,长期兴趣推荐方法具体效果如表 8 所示:

表 8 第一类长期兴趣推荐方法中各个用户相似度阈值限定下的平均推荐效果

用户相似度阈值范围	受测用户总数(单位:位)	匹配用户总数(单位:位)	平均推荐命中度
[90%,100%]	127	2 917	82.91%
[80%,90%)	53	4 558	82.71%
[70%,80%)	73	5 670	72.99%
[60%,70%)	85	6 907	66.15%
[50%,60%)	107	9 429	65.33%
[40%,50%)	118	12 075	63.45%
[30%,40%)	129	15 045	60.21%
[20%,30%)	131	18 214	56.78%
[10%,20%)	134	21 568	54.63%
[0%,10%)	139	24 407	53.31%

短期兴趣的即时推荐方法具体效果如表 9 所示:

表 9 第一类短期兴趣即时推荐方法中各个用户相似度阈值限定下的平均推荐效果

用户相似度阈值范围	受测用户总数(单位:位)	匹配用户总数(单位:位)	平均推荐命中度
[90%,100%]	103	2 673	82.47%
[80%,90%)	50	1 907	86.54%
[70%,80%)	62	1 654	70.07%
[60%,70%)	62	1 892	64.37%
[50%,60%)	76	2 384	59.83%
[40%,50%)	83	3 910	68.72%
[30%,40%)	91	3 728	66.27%
[20%,30%)	92	4 829	63.05%
[10%,20%)	97	5 698	54.86%
[0%,10%)	92	8 562	47.43%

对于第二类借阅量大于 100 条并且小于 200 条的读者用户,长期兴趣推荐方法具体效果如表 10 所示:

表10　第二类长期兴趣推荐方法中各个用户相似度阈值限定下的平均推荐效果

用户相似度阈值范围	受测用户总数（单位:位）	匹配用户总数（单位:位）	平均推荐命中度
[90%,100%]	1 038	4 337	84.99 %
[80%,90%)	385	5 509	79.24%
[70%,80%)	545	7 718	74.13%
[60%,70%)	701	9 991	71.74%
[50%,60%)	834	12 646	68.48%
[40%,50%)	972	15 521	68.58%
[30%,40%)	1 076	18 592	66.80%
[20%,30%)	1 134	21 510	63.45%
[10%,20%)	1 177	23 609	59.20%
[0%,10%)	1 209	24 491	55.71%

短期兴趣的即时推荐方法具体效果如表11所示：

表11　第二类短期兴趣即时推荐方法中各个用户相似度阈值限定下的平均推荐效果

用户相似度阈值范围	受测用户总数（单位:位）	匹配用户总数（单位:位）	平均推荐命中度
[90%,100%]	880	4 920	87.56%
[80%,90%)	384	4 168	82.94%
[70%,80%)	472	5 445	81.90%
[60%,70%)	536	5 631	78.11%
[50%,60%)	624	7 222	75.52%
[40%,50%)	680	7 971	74.08%
[30%,40%)	700	8 672	71.12%
[20%,30%)	748	9 418	65.01%
[10%,20%)	773	9 991	58.53%
[0%,10%)	718	9 659	49.58%

全部数据的相关对比如图2所示：

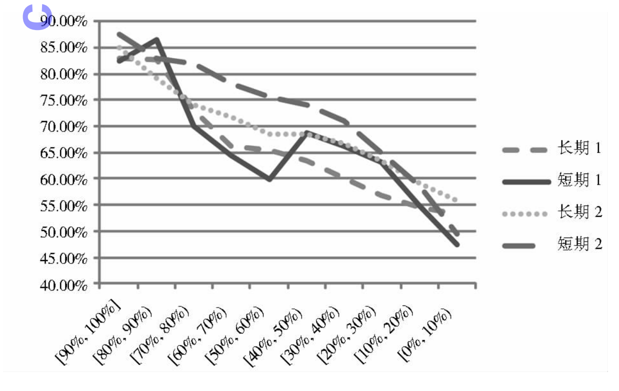


图2　两类读者用户的长期兴趣推荐和短期兴趣推荐的效果比较

从图2可以看出以下3个特点：

（1）不论长期兴趣推荐方法还是短期兴趣推荐方法，第二类用户的平均推荐命中度要高于第一类用户，主要原因在于第二类受测用户总数较高（为第一类用户的8.73倍），因此发现相似用户的概率更大，从而获取的有效推荐结果更多。但是在用户相似度位于75%以上近似区间内，不论长期方法还是短期方法，第一类用户平均推荐命中度反而要高于第二类用户。这在一定程度上说明，对于借阅量较高的读者用户而言，利用较高的用户相似度进行推荐具有优势。

（2）从总体看，在第一类用户中，最高相似度区间（75%以上）和较低相似度区间（15%到50%）的短期兴趣推荐方法的平均推荐命中度要高于长期兴趣推荐方法。在第二类用户中，绝大多数相似度区间（20%以上）的短期兴趣推荐方法的平均推荐命中度高于长期兴趣推荐方法，同时这也是所有平均推荐命中度最高的区间。从这里，可以看出短期兴趣推荐方法具有较为明显的优势，在较大的相似区间范围内都具有较为明显的平均推荐命中度，这也充分说明读者用户的阅读兴趣具有较为明显的时效性和短期特征，尤其是较近时期内的兴趣变化会对当前用户兴趣产生较大的影响。

（3）不论长期兴趣推荐方法还是短期兴趣推荐方法，都呈现出推荐命中度与用户相似度阈值范围的依赖关系，即放松用户相似度匹配范围，可以增加推荐用户数量，但是会给推荐命中度带来不利影响。

为了进行有效性对比，我们利用标准协同过滤方法对上述两类用户的短期兴趣推荐方法进行了测试，具体结果如表12所示：

表12　采用标准协调过滤方法实现的两类用户短期兴趣即时推荐方法结果

用户相似度阈值范围	第一类用户平均推荐命中度	第二类用户平均推荐命中度
[90%,100%]	2.03%	1.41%
[80%,90%)	2.23%	1.97%
[70%,80%)	2.3%	1.92%
[60%,70%)	2.17%	1.84%
[50%,60%)	2.4%	1.86%
[40%,50%)	2.28%	2.24%
[30%,40%)	2.31%	1.99%
[20%,30%)	2.4%	2.30%
[10%,20%)	1.89%	2.27%
[0%,10%)	1.76%	1.42%

从表12可以看出，两种方法总体推荐命中情况都不理想，其中第二类用户由于借阅量较小，推荐命中效果更差一些。这说明传统标准推荐方法由于没有考虑借阅历史所反映的用户兴趣演变趋势，同时也限于数据量有限，单纯使用相似度拟合的方法效果并不理想。

需要说明的是,目前的实验只是针对现有读者借阅记录,尚未进行优质图书资源推荐结合的考虑。为此,我们完成了一个用户在线满意度测试系统,邀请用户对自己感兴趣的图书进行检索,同时在南京财经大学图书馆借阅系统中开始进行相关实际推荐效果的用户测试,这部分工作需要一定的时间,待全部工作完成后,我们会对相关调研结果做出更为详细的分析和说明。

6　总结

作为对于图书借阅领域个性化推荐服务研究的一种探索,本文力图通过利用现有读者借阅记录得到的高质量专业性图书资源识别和读者用户相似度比较方法,提出改善现有图书馆个性化服务的策略,并进行实践,初步实现了预期的设计目标。不过,该方法仍然存在着需要进一步研究和改进的地方,主要问题在于目前利用读者借阅记录实现的个性化推荐必须建立在较大规模的读者用户借阅记录基础上,而且要求读者具有较为稳定的长期借阅习惯,反之对于那些偶尔借阅的读者用户而言,很难从借阅记录中得到更多的兴趣特征,不过这也能说明对这些用户而言,如何激发其关注图书借阅,对于改善图书馆图书借阅服务而言尤为重要。这构成了我们下一步的研究目标。

参考文献:

[1] WU F, HU Y, WANG P. Developing a novel recommender network-based ranking mechanism for library book acquisition[J]. Electronic library, 2017, 35(1):50 – 68.
[2] OUNI A, KULA R, KESSENTINI M, et al. Search-based software library recommendation using multi-objective optimization[J]. Information &software technology, 2017, 83(3):55 – 75.
[3] MATHEW P, KURIAKOSE B, HEGDE V. Book recommendation system through content based and collaborative filtering method[C]//Proceedings of international conference on data mining and advanced computing. Ernakulam, India:IEEE, 2016:47 – 52.
[4] 徐宾. 图书馆图书 h 指数的研究[J]. 情报学报,2014(8):892 – 896.
[5] 李克潮,梁正友. 基于多特征的个性化图书推荐算法[J]. 计算机工程,2012,38(11):34 – 37.
[6] PING H. The research on personalized recommendation algorithm of library based on big data and association rules[J]. Open cybernetics &systemics journal, 2015, 9(1):2554 – 2558.
[7] 凌霄娥, 周兵, 李克潮. 面向新读者和新图书的数字图书馆个性推荐冷启动问题研究[J]. 情报理论与实践, 2014, 37(8):100 – 104.
[8] 刘丹. 利用 Apache Mahout 部署个性化图书推荐服务[J]. 现代图书情报技术,2015,31(10):102 – 108.
[9] HE B, ZHANG H. Library personalized information recommendation of big data[C]//. Proceedings of online analysis and computing science. Chongqing, China:IEEE, 2016:289 – 292.
[10] 郑祥云,陈志刚, 黄瑞,等. 基于主题模型的个性化图书推荐算法[J]. 计算机应用,2015,35(9):2569 – 2573.
[11] 朱文奇. 推荐系统用户相似度计算方法研究[D]. 重庆:重庆大学, 2014.
[12] 马健,杜泽宇,李树青. 基于多兴趣特征分析的图书馆个性化图书推荐方法[J]. 现代图书情报技术,2012,28(6):1 – 8.
[13] 李克潮, 蓝冬梅, 凌霄娥. 云模型和多特征的高校读者借阅偏好不确定性图书推荐研究[J]. 现代图书情报技术,2013,29(5):54 – 58.
[14] 景民昌,于迎辉. 基于借阅时间评分的协同图书推荐模型与应用[J]. 图书情报工作,2012,56(3):117 – 120.
[15] 江周峰,鄂海红,杨俊. 基于时间上下文信息的借阅次数评分模型与应用[J]. 图书情报工作,2014,58(s2):220 – 223.
[16] 王进良,张鹏,狄增如,等. 北京师范大学图书借阅系统的网络分析[J]. 情报学报,2009,28(1):137 – 141.
[17] 傅林华,郭建峰,朱建阳. 图书馆图书借阅系统与单标度二元网络模型[J]. 情报学报,2004,23(5):571 – 575.
[18] 燕飞,张铭,孙韬,等. 基于网络特征的用户图书借阅行为分析——以北京大学图书馆为例[J]. 情报学报,2011,30(8):875 – 882.
[19] ZHAO S, ZHAO Y, SUN F, et al. Study on single mode weighted network of library lending network[J]. Journal of residuals science & technology, 2016, 13(6):241 – 246.
[20] 李树青,徐侠,许敏佳. 基于读者借阅二分网络的图书可推荐质量测度方法及个性化图书推荐服务[J]. 中国图书馆学报,2013,39(3):83 – 95.
[21] 袁虎声,赵洗尘. 基于加权借阅网络的个性化推荐算法与实现[J]. 图书情报工作,2016,60(10):130 – 134.
[22] 蓝冬梅. 大数据量图书下多数据集的二部图多样化推荐[J]. 情报理论与实践,2016,39(2):69 – 72.
[23] SUN Y, YIN H, REN X. Recommendation in context-rich environment: an information network analysis approach[C]//Proceedings of the 26th international conference on World Wide Web companion. Perth, Australia: International WWWconferences steering committee, 2017:941 – 945.
[24] 邱均平,张聪. 高校图书馆馆藏资源协同推荐系统研究[J]. 图书情报工作,2013,57(22):132 – 137.
[25] 李树青,孙颖. 基于加权关键词共现时间元的个性化学术研究时序路径发现及其可视化呈现方法[J]. 情报学报,2014,33(1):55 – 67.
[26] PAGE L. The PageRank citation ranking : bringing order to the web[J]. Stanford Digital Libraries Working Paper, 1998, 9(1):1 – 14.

作者贡献说明:

李树青:提出研究思路,进行论文撰写、修改及最终版

本修订；
庄光光:进行实验算法设计和相关数据整理分析；
秦嘉杭:负责原始借阅记录信息整理和相关数据处理；
徐侠:负责推荐实验结果评价分析。

The Method of Measuring the Professional Quality of Books and Personalized Book Recommendation Service in Circulating Scene

Li Shuqing[1] Zhuang Guangguang[1] Qin Jiahang[2] Xu Xia[3]
[1] School of Information Engineering, Nanjing University of Finance and Economics, Nanjing 210046
[2] Library, Nanjing University of Finance and Economics, Nanjing 210046
[3] School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210046

Abstract: [**Purpose/significance**] With the analysis of the existing library records and mining of the reading relevance of books, this paper discusses the effective methods to identify high-quality professional books and implement a personalized recommendation service. [**Method/process**] This paper introduces the iterative algorithm of recognizing high-quality professional books from links of books relevance based on reading relevance. Then the construction of reader personalized profile is discussed based on the definition of links of book categories. The design and implementation of long-term and short-term personalized recommendation methods are also proposed. [**Result/conclusion**] In the aspect of book quality identification, it is easier to identify the professional books resources with higher quality. This application is more flexible and also can identify the high-quality professional books within the collection of specific books. It is found that whether long-term or short-term interest interest recommendation method, the average hit degree of users with higher lending is higher than users with lower lending. In the group of users with higher lending, the average hit degree of short-term interest recommendation method in the highest similarity range (more than 75%) and lower similarity range (15% to 50%) is higher than the long-term interest recommendation method.

Keywords: personalized recommendation book borrowing library service book quality